Editors' Introduction

Sanjay Jain, Rémi Munos, Frank Stephan, and Thomas Zeugmann

The aim of the series of conferences on Algorithmic Learning Theory (ALT) is to look at learning from algorithmic and mathematical perspective. Over time several models of learning have been developed which study different aspects of learning. In the following we describe in brief the invited talks and the contributed papers for ALT 2013 held in Singapore.

Invited Talks. Following the tradition of the co-located conferences ALT and DS all invited lectures are shared by the two conferences. The invited speakers are eminent researchers in their fields and present either their specific research area or lecture about a topic of broader interest.

This year's joint invited speaker for ALT 2013 and DS 2013 is Nir Ailon, who is an assistant professor at the Technion Department of Computer Science since 2010. He received his Ph.D. from Princeton University in 2006. Then he joined the Institute for Advanced Study in Princeton for a year as a postdoctoral member. Subsequently, he worked at Google Research for two years. His research interests comprise high dimensional statistics, dimensionality reduction techniques, learning theory, preference analysis, and ranking and clustering. His research contributions have been recognized by a SIAM outstanding paper award in 2012. The invited talk *Learning and Optimizing with Preferences* by Nir Ailon deals with reasoning about preferences and choices as a fundamental source of information. The social sciences have studied preferences and choices for centuries and with the advent of the internet, preference and ranking data became ubiquitous. This abundance attracted many computer scientists to study such data in different contexts such as in information retrieval and filtering, rank aggregation, learning theory, among others. Ailon surveys fundamental results, points to important problems, and exemplifies the challenges that arise in this context. In particular, he points out that preferential information can be used to optimize information systems.

The invited speaker for ALT 2013 is Eiji Takimoto, who is a professor at the Department of Informatics at Kyushu University, Fukuoka, Japan. His major research interests are online decision making, computational learning theory, and complexity theory. The invited paper *Efficient Algorithms for Combinatorial Online Prediction* (co-authored by Kohei Hatano) deals with the following problem. Let a concept class $\mathcal{C}$ be given which is typically finite but contains exponentially many concepts. Each concept of the class is encoded as an n-dimensional non-negative real valued vector, where n is a natural number. Furthermore, a so-called loss space $\mathcal{L}$ is given which consists also of non-negative real valued vectors. The online linear optimization problem specified by a concept class $\mathcal{C}$ and the loss space $\mathcal{L}$ can then be described as a repeated game between the player and the adversary. For each trial $t = 1, 2, \ldots, T$, the player chooses a concept

$c_t \in \mathcal{C}$ and the adversary returns a loss vector $\ell_t \in \mathcal{L}$. The player suffers a loss $c_t \cdot \ell_t$. The task is to minimize the expected regret. There are many interesting problems that can be studied in this model such as spanning trees of a given graph, truth assignments for a given CNF formula, and so on. The paper surveys some recent results on universal and efficient implementations of low-regret algorithmic frameworks such as Follow the Regularized Leader (FTRL) and Follow the Perturbed Leader (FPL).

Nader H. Bshouty is the ALT 2013 tutorial speaker. He is the Helen and Morris Mauerberger Chair in Sciences at the Technion Department of Computer Science. His major research interest is computational learning theory. In his talk *Exact Learning from Membership Queries: Some Techniques, Results and New Directions* he focuses on the following general learning scenario. We are given a black box that contains a function $f\colon D \to R$ from some class $\mathcal{C}$ of functions. The source of information for the learner are membership queries, i.e., the learner chooses in each time step an element d from the domain D of the function f, and then the black box returns $f(d) \in R$ in time T. The learning goal is then to exactly learn the function f with a minimum number of queries and optimal time complexity. Sometimes also a weaker learning goal is considered. That is, the learner has to decide whether or not the function f is equivalent to a prespecified function $g \in \mathcal{C}$. Such learning problems have been studied in different areas and also under a variety of names such as interpolation, active learning, inference, guessing game, functional verification, and so on. Bshouty provides an extensive survey of the results obtained and outlines new directions that are worth investigating.

Hannu Toivonen is the invited speaker for DS 2013. Since 2002, he is Professor of Computer Science at the University of Helsinki, Finland. Prior to the current position, he worked at Nokia Research Center as principal scientist. His research interests comprise data mining, computational method for data analysis, and applications of these areas in bioinformatics, genetics, ecology, and mobile communications. Currently, he and his research group focus their interest to computational creativity and this area is addressed in his talk *Creative Computers and Data Mining*. A major goal in computational creativity is to enable computers to perform tasks that require creative skills such as those needed in writing poetry or composing music. Clearly, a computational agent then needs to know the field in which it operates. Toivonen and his research group discovered that data mining has a great potential for achieving this goal. He outlines how to make creative agents adaptive to various fields and genres by automatic discovery of relevant information from existing creative artifacts. The talk gives several examples of how verbal creativity can benefit from data mining of existing text corpora. Furthermore, he exemplifies that computational creativity tools allow a whole new approach to data analysis. In this "Affective Data Analysis," the goal is to turn data into a subjective, esthetic experience by automatic or semiautomatic creation of a novel artifact using the user's data as inspiration. This idea is illustrated with musicalization of sleep measurements and chat discussions.

Online Learning. In online learning, a problem (or a component of it) has to be solved online. At time t, some instance of a problem (or part thereof) is given to the learner, and it outputs a solution (or part thereof). The online algorithm suffers a regret based on how badly it performs compared to the best offline algorithm, which knows the whole input. In some cases, the offline algorithms might be required to have some properties, such as being deterministic, whereas the online algorithm might be probabilistic.

One of the uses of online prediction is in stock market, where one may want to predict as well as any other strategy. The paper *Universal Algorithm for Trading in Stock Market Based on the Method of Calibration* by Vyugin considers such a problem. Suppose $S_1, S_2, \ldots$, are prices of a stock over time. Assume via scaling that each S_i is in the range $[0, 1]$. Suppose a trader trades in each interval, and gains (looses) based on the change in price. Furthermore, the paper assumes a side information $z_i \in [0, 1]$, a signal, is given to the trader. Consider any stationary trading strategy, D, which is a continuous function of z_i, that is, the trader buys/sells $D(z_i)$ units of share at time i. Then Vyugin constructs a randomizing algorithm M such that, $\liminf_{n\to\infty} \frac{1}{n}(K_n^M - \|D\|_+^{-1} K_n^D) \geq 0$, almost surely, where K_n^M and K_n^D are the gain made by M and D respectively after n time steps, and $\|D\|_+^{-1}$ is a normalizing factor used (as D may buy/sell more than one units, though M is restricted to buy/sell 1 unit in each time step).

Online prediction of combinatorial concepts such as set cover, permutations, MAX-SAT etc arise in real life situations such as in routing and ranking. Suppose $\mathcal{C} \subseteq \mathbb{R}^n$ is a finite set of combinatorial concepts. Consider the following protocol for online prediction: in each trial t, the player predicts $c_t \in \mathcal{C}$, the adversary returns loss vector ℓ_t, and the player has a loss of $c_t \cdot \ell_t$. The aim of the player is to minimize the loss compared to the best possible $c \in \mathcal{C}$ that is, $\sum_t c_t \cdot \ell_t - \min_{c\in\mathcal{C}} \sum_t c \cdot \ell_t$. Fujita, Hatano and Takimoto in the paper *Combinatorial Online Prediction via Metarounding* consider a variant in which one minimizes the α regret, that is $\sum_t c_t \cdot \ell_t - \alpha \min_{c\in\mathcal{C}} \sum_t c \cdot \ell_t$. This is based on the fact that for many combinatorial problems, α-approximation algorithms are known, which are allowed to be used as oracles by the online algorithm. Earlier, Kalai and Vempala had proposed a *Follow the Perturbed Leader* (FPL) algorithm, which works for $\alpha = 1$. This algorithm has a regret bounded by $O(\sqrt{T})$ and runs in time $O(n)$ per trial. Taking into account α-approximations, this algorithm has a regret $O(\alpha^T\sqrt{T})$, which is high. Kakade, Kalai and Ligett proposed another strategy which achieves $O(\alpha\sqrt{T})$ regret with running time of $O(poly(n)T)$. The paper by Fujita, Hatano and Takimoto considers a stronger assumption on an approximation algorithm: for any $\ell \in [0, 1]^n$ as input, the approximation algorithm outputs a $c \in \mathcal{C}$ such that $c \cdot \ell \leq \alpha \min_{x\in\mathcal{P}} x \cdot \ell$, where $\mathcal{P}$ is a convex superset of $\mathcal{C}$ and linear optimization over $\mathcal{P}$ can be done in time polynomial in n. This is based on the fact that several combinatorial problems have corresponding linear programming approximations. Under this assumption, Fujita, Hatano and Takimoto give a strategy whose regret is bounded by $O((\alpha+\varepsilon)\sqrt{T})$ and has running time polynomial in n and $1/\varepsilon$ for any $\varepsilon > 0$. The main idea

used is the concept of metarounding by boosting. Metarounding was originally proposed by Carr and Vempala for a different purpose. The authors show as examples that the method works for online set cover, MAX-SAT, and some other combinatorial problems.

In their paper *On Competitive Recommendations* Uitto and Wattenhofer consider the problem of recommending an item to a user. Suppose M is an unknown $n \times m$ binary matrix. The goal of the algorithm is to determine at least one 1 entry in each row, using a minimum number of queries, where it is assumed that each row contains at least one 1. The entries can be considered as preferences of n users (customers) on m items (say books): $M(i,j) = 1$ can be considered as customer i liking book j. The goal of the algorithm is to find a suitable book for each customer. To determine a suitable book for customer i, the algorithm can query an entry of the matrix. The protocol of the algorithm works in rounds, where each round has three steps, (where U is the set of unsatisfied customers, initially the set of all the customers): (1) receive a customer $u \in U$ chosen uniformly at random, (2) recommend book b to u, and (3) if $M(u,b) = 1$, then the algorithm can remove u from U. One measures the number of rounds needed (equivalently the number of recommendations made in the step (2) above). The cost can thus vary from n to $m \cdot n$. Comparing an on-line algorithm against an off-line algorithm which knows all the entries of the matrix is not much useful, as the off-line algorithm can then solve the problem using n queries. Thus, Uitto and Wattenhofer consider the competitiveness against a quasi-off-line algorithm which only knows the probability distribution D over possible preference vectors and n, the number of customers. The preference vectors for these customers are chosen independently at random from D. The authors consider the case of $m = O(n)$, as if m is very large, many of columns might have 0 for every customer. Uitto and Wattenhofer give a $O(\sqrt{n}\log^2 n)$ competitive on-line algorithm for the problem. The authors also show a lower bound of $\Omega(\sqrt{n})$ for the above problem, and thus their algorithm is within a poly-log factor of optimal competitive ratio.

The paper *On-line PCA with Optimal Regrets* by Nie, Kotłowski and Warmuth considers the online version of the Principal Component Analysis (PCA) problem. In PCA, an n-dimensional datum is given as input, and the output is a *projection* of it in k-dimensions, where k is much smaller than n. In their paper, the following model is considered. In each trial $t = 1, 2, \ldots, T$ the algorithm first chooses a projection matrix $\mathbf{P}_t$ of rank k. Then it is given the next point $\mathbf{x}_t$ of dimension n. The algorithm suffers a compression loss $\|\mathbf{x}_t - \mathbf{P}_t\mathbf{x}_t\|_2^2 = tr((\mathbf{I} - \mathbf{P}_t)\mathbf{x}_t\mathbf{x}_t^T$. The goal is to obtain an on-line algorithm whose cumulative loss is not far from the loss of the best possible projection matrix $\mathbf{P}$ which is chosen after having seen all the data. There are two main types of algorithms which have been used for on-line learning. The Gradient Descent (GD) family of algorithms and Exponentiated Gradient (EG) family of algorithms. For PCA, the authors show that EG achieves the same regret bound as GD despite the sparseness of instance matrices (for which usually EG algo-

rithms do not do well). They also show that for both algorithms, regret bounds are within a constant factor of a lower bound for any PCA algorithm.

Inductive Inference and Grammatical Inference. The learning model studied in inductive inference of formal languages from positive data may be described as follows. A learner receives, one element at a time, all the elements of a target language L from a class of languages $\mathcal{L}$. Over time, each element of the target language is presented at least once (in arbitrary order) to the learner and no non-elements of the language L are presented to the learner. As the learner receives its data, it conjectures a sequence of hypotheses. These hypotheses are interpreted as grammars in some system which contains grammars for all the languages in $\mathcal{L}$. If the sequence of grammars converges to a grammar for L, then one can say that the learner has learned the language L. The learner learns $\mathcal{L}$ if it learns each language in $\mathcal{L}$. This is basically the model of *explanatory* learning first considered by Gold, and later by various authors. Over time various modifications to the above model have been considered. A variation of the model of explanatory learning, first studied by Osherson, Stob and Weinstein, is *partial learning*, where the learner outputs exactly one index infinitely often, and this index is for the input language. In the traditional explanatory learning model, various researchers have considered constraints on the learner such as (i) consistency (first introduced by Bārzdiņš), where each conjecture of the learner is expected to be consistent with the data seen by the time the conjecture is made, i.e., the conjectured language must contain the data seen so far, (ii) conservativeness (first studied by Angluin), where the learner changes its hypothesis to a different one only if the presented data are inconsistent with its hypothesis, and (iii) confidence, where the sequence of hypotheses of the learner converges to a hypothesis, whether correct or wrong, even when the data presented to it is for a language outside the class being learned.

In the paper *Partial Learning of Recursively Enumerable Languages*, Gao, Stephan and Zilles study the corresponding notions for partial learning. They give a complete picture of relationships between these criteria. Furthermore, the authors also give interesting characterizations of some of the inference criteria studied. In particular they show that a class is consistently partially learnable iff it is a subclass of a uniformly recursive family of languages.

One of the main problems considered in inductive inference is comparison between different learning criteria. That is, given two learning criteria I and J, is every class of languages (functions) learnable under criterion I also learnable under criterion J? There are two main reasons for a class not to be learnable under a criterion: *topological* reasons, when the class is not learnable under the criterion even if one allows the learners to be non-computable; and *computational* reasons, where the class is not learnable for computable learners, but would become learnable if the computability constraints are removed from the learner. The paper *Topological Separations in Inductive Inference* by Case and Kötzing formalizes when two learning criteria separate topologically in learning power. This allows one to study more closely the relative powers of the two criteria. For example, if one considers the model $TxtEx^a$, where in the explanatory learning

model described above one allows the final conjecture to have upto a errors, then one can show that $TxtEx^{a+1}$ and $TxtEx^{a}$ separate topologically. However, if one considers vacillatory learning, that is the number of different grammars output by the learner during the learning process is finite, and after some time, all the grammar output by the learner are correct, then it can be shown that vacillatory learning separates from explanatory learning, but not topologically. Case and Kötzing show for variety of pairs of learning criteria from the literature that they separate topologically, whereas some pairs of criteria are shown not to separate topologically.

The class of all recursively enumerable languages or even that of all context-free languages or of all regular languages is not learnable under various models of learning, such as explanatory learning. Thus it is interesting to consider special subclasses of context-free languages which are learnable, particularly if they can be learned efficiently. In the literature there have been several results showing interesting subclasses of context free languages to be learnable under the explanatory learning model and/or other learning models. Some examples include explanatory learning of substitutable context free languages and learning of c-deterministic and congruential context free languages from a minimally adequate teacher. However, these results are often impractical, as one may never know when the learner has converged to the correct grammar in explanatory learning and a minimally adequate teacher is often not available. Probabilistic learning is often considered more practical. However, the positive results for learning such classes in Valiant's PAC learning model are rather limited. Clark has shown that a class of unambiguous non-terminally separated languages is PAC learnable from partially distribution-free positive data. Luque and Lopez later generalized this result. The paper *PAC Learning of Some Subclasses of Context-Free Grammars with Basic Distributional Properties* by Shibata and Yoshinaka expands Clark's result to other subclasses of context free grammars that are known to be exactly learnable based on distributional learning techniques. Shibata and Yoshinaka translate some of the existing distributional exact learning algorithms into PAC-type ones where a learner gets positive examples drawn from a distribution determined by a probabilistic context free grammar. Under some assumptions, Shibata and Yoshinaka show how membership queries used in exact learners can be simulated by observed positive examples.

A goal of a scientist is to acquire knowledge about the surrounding and to predict the future. The Solomonoff induction method solves the prediction problem by using algorithmic information theory to obtain a universal prior and Bayes theorem to perform induction. This method is able to predict in any stochastically computable environment. However, the method itself is uncomputable. Solomonoff induction has been extended to the reinforcement learning framework by Hutter. However, optimal reinforcement learning is different from an optimal scientist as it is rewarded extrinsically by the environment rather than intrinsically by the information gain. A knowledge seeking scientist tries to maximize the information gain. Storck, Hochreiter, and Schmidhuber used various information gain criteria in a frequentist setting to explore non-deterministic

Markov environments, where information gain is considered as rewards. Orseau had earlier presented two universal knowledge seeking agents, called Square-KSA and Shannon-KSA for deterministic computable environments. In the paper *Universal Knowledge-Seeking Agents for Stochastic Environments*, Orseau, Lattimore, and Hutter consider countable stochastic environments, and define a new universal knowledge-seeking agent based on Kullback-Leibler divergence. The resulting universal agent, if it exists, has several nice properties. The new agent is resistant to noise and behaves as expected in variety of toy examples.

Teaching and Learning from Queries. Suppose $X = \{x_1, x_2, \ldots, x_n\}$ to be an instance space and assume that $\mathcal{C}$ and $\mathcal{H}$ are subsets of the powerset of X. Then $\mathcal{C}$ and $\mathcal{H}$ are said to be a concept class and a hypothesis class, respectively. A labeled sample $S \subseteq X \times \{0,1\}$ is $\mathcal{C}$ realizable if there exists a $C \in \mathcal{C}$ such that, for all $(x,\ell) \in S$, $C(x) = \ell$. A sample compression scheme, for a given class $\mathcal{C}$, consists of a compression function f and a reconstruction function g, with the property that, for any labeled sample S, the conditions $f(S) \subseteq S$ and $g(f(S))(x) = \ell$ are satisfied for all $(x,\ell) \in S$. An open question is whether for any class $\mathcal{C}$, there exists a sample compression scheme of size linear (or even equal) to the VC-dimension of $\mathcal{C}$. A *teaching set* for a concept $C \in \mathcal{C}$ is a sample S such that C is the only concept in $\mathcal{C}$ which is consistent with S. A *teaching plan* for the (ordered) class $\mathcal{C} = \{C_1, C_2, \ldots, C_m\}$ is a sequence of samples $((C_1, S_1), (C_2, S_2), \ldots, (C_m, S_m))$ such that S_t is a teaching set for C_t with respect to the class $\{C_t, C_{t+1}, \ldots, C_m\}$. The order of the teaching plan is the maximum over the cardinality of the S_i's. The *recursive teaching dimension* (RTD) of $\mathcal{C}$ is the minimum order over all teaching plans for $\mathcal{C}$. The paper *Order Compression Schemes* by Darnstädt, Doliwa, Simon, and Zilles considers a sample compression scheme called order compression scheme. In this, for the hypothesis class $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$, $f(S)$ is the smallest subset of S that is a teaching set for H_t with respect to $\{H_t, H_{t+1}, \ldots, H_m\}$, where t is the largest number such that H_t is consistent with S. Furthermore, $g(f(S)) = H_r$, for the largest number r such that H_r is consistent with $f(S)$. The authors show that such a scheme is indeed a compression scheme. Let $\mathrm{OCN}(\mathcal{C}, \mathcal{H})$ denote the cardinality of the largest compressed sample in the order compression scheme defined above, and $\mathrm{OCN}(\mathcal{C})$ denote the minimal $\mathrm{OCN}(\mathcal{C}, \mathcal{H})$ over all $\mathcal{H} \supseteq \mathcal{C}$. Then the authors show that $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{RTD}(\mathcal{C})$, and $\mathrm{OCN}(\mathcal{C}) \geq \mathrm{VCD}(\mathcal{C})$. Furthermore, for many natural classes, such as those which are intersection-closed or maximum or Dudley classes, the authors show that $\mathrm{OCN}(\mathcal{C}) = \mathrm{VCD}(\mathcal{C})$. Thus, order compression schemes give a reasonable way to try to address the sample compression conjecture.

In the paper *Learning a Bounded-Degree Tree using Separator Queries*, Jagadish and Sen investigate the following problem. Consider an undirected tree T with bounded degree. Suppose the learner knows the nodes in the tree (numbered 1 to n), but does not know the edges of the tree. The learner has to find the tree T using the following type of queries called separator queries: Does node x lie in the path from node a to node b? Jagadish and Sen show that there exists an $O(n^{1.5} d \log n)$ time algorithm for the above problem, where d is the bound

on the degree of the tree, and n is the number of nodes of the tree. They also give an $O(nd^2 \log^2 n)$ randomized algorithm and show a lower bound of $\Omega(nd)$ for the problem.

Bandit Theory. In multi-armed bandit problems, the standard stochastic setting considers an agent or learner facing a finite number of distributions (also called arms), that can be sampled one at a time. Each sample gives a reward, and the goal is to maximize the reward after T trials. As the learner/agent's decision is made based on random data from an unknown distribution, there is often a tradeoff between exploration (drawing more arms) and exploitation (drawing the current best arm). For the analysis, the learner is often compared to a fixed strategy which constantly pulls the arm with highest mean reward. The expected regret is then the difference between the learner's cumulative reward and the cumulative reward of the fixed strategy as described above.

The paper *Faster Hoeffding Racing: Bernstein Races via Jackknife Estimates* by Loh and Nowozin considers a variation of the bandit problem where the goal is not to minimize the expected regret but to identify an almost optimal arm with high confidence using as few resources as possible. That is, the aim is to determine how many samples to use before concluding that with probability at least $1 - \delta$, the selected arm is within a factor $1 - \varepsilon$ of optimal. Often racing algorithms are used, where confidence intervals are constructed. As the aim is to get a (near) optimal arm, one eliminates arms with low value after few samples, and then races the best candidates against one another. Hoeffding race derives its name from Hoeffding's inequality, which is used to construct the confidence intervals. Loh and Nowozin use tighter variants based on empirical Bernstein inequalities as well as jackknife estimates for constructing confidence intervals. They show that this gives better bounds for U-statistics and discrete entropy estimators.

The usual situation in bandit problems is for the learner to maximize its expected return (or minimize the expected regret). However this may not be suitable for every problem. For example, in medical treatment, one may want to avoid endangering the patient. So one wants to avoid high variation in the rewards. In particular, one may prefer an arm with smaller mean, but narrower left-tail compared to an arm with larger mean, but fat left-tail. Maillard's paper *Robust Risk-averse Stochastic Multi-Armed Bandits* addresses such a situation. This paper defines a notion of risk-averseness based on the best risk-averseness of the arms, and then gives an algorithm, called RA-UCB, which has $O(\log T)$ bounds on the regret and $O(\log T)$ bounds on the risk-averse regret, with respect to the best risk-averse arm, where T is the number of trials.

The paper *An Efficient Algorithm for Learning with Semi-Bandit Feedback* by Neu and Bartók considers a semi-bandit setting, where one can pull several arms together and see the loss/reward for each of the arms pulled. Only certain combinations of arms are allowed to be pulled together. An example where this could be useful is displaying m advertisements out of a possible n advertisements when a user visits a web-page. The authors propose an algorithm combining the Follow-the-Perturbed-Leader (FPL) prediction method with a new

loss estimation procedure which they call Geometric Resampling (GR). This new algorithm can be efficiently implemented whenever efficient offline combinatorial optimization is possible. The authors show the expected regret, after T rounds, to be $O(m\sqrt{dT \log d})$, where the elements of the decision set is described using d-dimensional binary vectors with at most m non-zero entries. This also improves the best known regret bounds for FPL in the full information setting to $O(m^{3/2}\sqrt{T \log d})$.

Statistical Learning Theory. Many real world datasets have a high dimension. The time-complexity of many existing algorithms dealing with a large amount of data often depends super-polynomially on the dimension. This is called the curse of dimensionality in machine learning. Often, the dimension of the data-sets can be reduced as the actual data lie in a low dimensional manifold. Several dimension reduction techniques have been developed. Dasgupta and Freund have analyzed the technique presented by Freund et al. to learn the structure of a manifold that has low dimension d, but the actual data lie in $\mathbb{R}^D$, where D is much larger than d. This involved a construction of a data structure called random projection tree. Choromanska, Choromanski, Jagannathan, and Monteleoni, in the paper *Differentially-Private Learning of Low Dimensional Manifolds*, extend this technique for the case of differential private learning. Differential privacy is a model of privacy for database access. The aim in this setting is to make sure that the addition/removal of a single data item in the database does not cause significant impact on the output produced for a query. The problems of high-dimensionality are compounded in differentially-private learning as it needs more data. The authors extend the random projection tree technique by constructing a differentially-private data structure that depends exponentially only on the doubling dimension d of the data rather than the extrinsic dimension D.

In traditional learning theory often the assumption is made that the data are independent and identically distributed (IID). However, this may not always hold in the real world. When training data come form a Markov chain with certain mixing properties, a common algorithm that has been analyzed is the empirical risk minimization (ERM) algorithm, which tries to find the hypothesis which minimizes the empirical loss. Generalization bounds of ERM have been proved for strongly mixing data, uniformly ergodic data, and V-geometrically ergodic data. The paper *Generalization and Robustness of Batched Weighted Average Algorithm with V-geometrically Ergodic Markov Data* by Cuong, Ho, and Dinh, considers another learning algorithm called the batched weighted average (BWA) algorithm. This algorithm places weights on each of the hypotheses during training. During testing a prediction is made based on the weighted average prediction of the hypotheses. The advantage of such a method is that it suffers less from overfitting. Cuong, Ho, and Dinh give a PAC-style bound on the training sample size for the expected loss to converge to optimal loss with high probability when the training data are V-geometrically ergodic. The authors also show robustness of BWA in the presence of a small amount of noise.

Most of the dimension reduction techniques require that the date lie on a low dimension subspace, and do not work when the sample is *close* to a low-dimensional subspace. The paper *Adaptive Metric Dimensionality Reduction* by Gottlieb, Kontorovich and Krauthgamer addresses this issue when the data are close to a low-dimensional subspace. The authors show that the linear classifier generalizes well for such data regardless of the ambient dimension. This result is shown for the Euclidean space and then generalized to metric spaces.

The paper *Dimension-Adaptive Bounds on Compressive FLD Classification* by Kabán and Durrant continues the work on how to address the curse of dimensionality by using the intrinsic dimension rather than the ambient dimension. This paper analyzes the Compressive Fisher Linear Discriminant (CFLD) classifier and shows that, under certain conditions, the estimated error of the CFLD depends on the intrinsic dimension of the data rather than the ambient dimension. The authors also show that if the data is noisy, then dependence on the ambient dimension cannot be totally avoided.

Bayesian/Stochastic Learning. In the paper *Bayesian Methods for Low-Rank Matrix Estimation: Short Survey and Theoretical Study*, Alquier studies low-rank matrix learning. Despite empirical performance of Bayesian procedures being at least as good as the performance based on penalized empirical risk minimization methods, there have not been many theoretical guarantees on Bayesian procedures. In his paper, Alquier gives a theoretical result on the Bayesian estimator in the context of reduced rank regression. For some appropriate choice of the parameters, the rate of convergence is the same as that for penalized methods, up to log terms.

The sequence prediction problem is to predict x_t, having already observed $x_1, x_2, \ldots, x_{t-1}$. The sequence is assumed to be sampled from an unknown measure μ contained in a countable model class $\mathcal{M}$. At time t, having observed $x_1, x_2, \ldots, x_{t-1}$, the predictor outputs a distribution ρ_t over the next symbol x_t. A predictor can be considered good if, $\rho_t - \mu_t \to 0$. Bayesian prediction methods are often considered, where one assigns a non-zero prior probability to each measure in $\mathcal{M}$. Let ξ denote the probability distribution obtained for a universal Bayesian predictor. Strong bounds are known on the expected cumulative error $\sum_t dist(\xi_t, \mu_t)$ with respect to the Kullback-Leibler divergence and Hellinger distance. The paper *Concentration and Confidence for Discrete Bayesian Sequence Predictors* by Lattimore, Hutter and Sunehag shows a high-probability bound on this cumulative error: with respect to the Kullback-Leibler divergence, the cumulative error is bounded by $e \cdot (\ln \frac{6}{\delta})(\ln \frac{2}{\delta} + ln \frac{1}{w_\mu})$, with μ-probability at least $1 - \delta$, where μ is the underlying probability distribution. The authors also show that this bound is close to optimal. Furthermore, the authors show that the Kullback-Leibler divergence $dist(\xi_t, \mu_t)$ can be bounded without knowing μ (but knowing a bound on prior probability for μ), with high confidence.

Convex optimization is the task of approximately minimizing a convex function over a convex set, given oracle access to unbiased estimates of the function and its gradient at any point, thereby using as few queries as possible. The problem of active threshold learning can be described as follows. Suppose we have

an interval $[0, R]$. A point $x \in [0, R]$ has label + and − with probability $\eta(x)$ and $1 - \eta(x)$, respectively. Assume that there exists a unique point t such that $\eta(t) = 1/2$, and $\eta(t) > 1/2$ on one side of t and $\eta(t) < 1/2$ on the other side of t. Then the task is to estimate t by sequentially querying T points and observing the labels from the distribution. Recently some connections have been established between convex optimization and active learning. In the paper *Algorithmic Connections Between Active Learning and Stochastic Convex Optimization*, Ramdas and Singh continue to exploit these connections. Inspired by a recent optimization algorithm that was adaptive to unknown uniform convexity parameters, the authors present a new active learning algorithm for one-dimensional thresholds that is adaptive to unknown noise parameters. Furthermore, the authors show that noisy gradient signs suffice for minimization of uniformly convex functions by showing that a random coordinate descent algorithm with an active learning line-search subroutine achieves minimax convergence rates.

Unsupervised/Semi-Supervised Learning. In many learning situations, the feedback for the learning algorithm is scarce or absent. In the paper, *Unsupervised Model-Free Representation Learning*, Ryabko considers a situation in which all or a large part of relevant information is in the time-series dependence of the process. Such a situation happens for example in speech or hand-written text or sensor data interacting with the environment. Assume that there is a stationary sequence $X_0, X_1, \ldots, X_n, \ldots$, where X_i belongs to a large continuous and high dimensional space $\mathcal{X}$. The aim is to look for a compact representation $f(X_0), f(X_1), \ldots$, where $f(X_i)$ belongs to a small space $\mathcal{Y}$. In an ideal situation, there exists a function f as above such that, given $f(X_i)$, X_i is independent of rest of the sample $X_0, X_1, \ldots, X_{i-1}, X_{i+1}, \ldots$. Thus all the time-series dependence of the sequence is available in $f(X_0), f(X_1), \ldots$, and given this sequence, the X_i are conditionally independent. In this ideal situation, it can be shown that f maximizes $I_\infty = h(f(X_0)) - h_\infty(f(X))$, where $h(f(X_0))$ gives the Shannon entropy of the first element and h_∞ is the entropy rate of the stationary time series $f(X_0), f(X_1), \ldots$. In a non-ideal situation one may define the function f which maximizes I_∞ as the one which preserves the most of the time-series dependence. In the paper, Ryabko shows that under certain conditions it is possible to estimate I_∞ uniformly over a set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathcal{Y}$. This can even be done without estimating the distribution of the original time series $(X_i)_{i\in\mathbb{N}}$. In particular, if the sequence $(X_i)_{i\in\mathbb{N}}$ forms a Markov process, then in the ideal situation, it can be shown that $(f(X_i))_{i\in\mathbb{N}}$ is also Markov and that $I_\infty(f) = h(f(X_0)) - h(f(X_1)|f(X_0))$. Rybko also extends the results to the case when the learner is allowed to take some actions (which may affect the next observations).

Clustering has become one of the fundamental problems in machine learning due to the presence of large datasets. Spectral clustering is one of the existing techniques for clustering. A problem for scaling up this method is the cost of building an affinity matrix between pairs of data points, which becomes computationally prohibitive for large data sets. There have been several attempts to address the problem, however, most of these works did not provide perfor-

mance guarantees. The paper *Fast Spectral Clustering via the Nyström Method* by Choromanska, Jebara, Kim, Mohan, and Monteleoni, gives a computationally efficient modification of the spectral clustering algorithm. The authors combine the spectral clustering idea with the Nyström approximation method. For this they only need to sample a small random set of columns of the affinity matrix, which leads to a complexity that is linear in the number of data points. The authors also provide performance guarantees for their algorithm, which are comparable to spectral clustering with the original graph Laplacian.

Often, the distribution from which data are sampled may change over time. Suppose we have a sample

$$x := X_1, X_2, \ldots, X_{\lfloor n\theta_1 \rfloor}, X_{\lfloor n\theta_1 \rfloor + 1}, \ldots, X_{\lfloor n\theta_2 \rfloor}, X_{\lfloor n\theta_2 \rfloor + 1}, \ldots, X_{\lfloor n\theta_\kappa \rfloor + 1}, X_n$$

formed by the concatenation of $\kappa + 1$ non-overlapping segments, where θ_1, $\theta_2, \ldots, \theta_\kappa \in (0, 1)$ are parameters. Each segment is generated by some unknown stochastic process distribution. The aim is then to determine the change points $\lfloor n\theta_1 \rfloor, \lfloor n\theta_2 \rfloor, \ldots, \lfloor n\theta_\kappa \rfloor$. In the paper, *Nonparametric Multiple Change Point Estimation in Highly Dependent Time Series*, Khaleghi and Ryabko consider such a problem. They consider a highly dependent time series, where each segment is generated by an unknown stationary ergodic process distribution. The joint distribution over the samples can be otherwise arbitrary. The authors construct an asymptotically consistent algorithm that estimates the parameters $\theta_1, \theta_2, \ldots, \theta_\kappa$, where the estimates becomes arbitrary close to the actual one as n goes to infinity.